ARMY RESEARCH LABORATORY

Feature Extraction of High-Dimensional Structures for Exploratory Analytics

by Andrew M. Neiderer

ARL-TN-531 **April 2013**

NOTICES

Disclaimers

The findings in this report are not to be construed as an official Department of the Army position unless so designated by other authorized documents.

Citation of manufacturer's or trade names does not constitute an official endorsement or approval of the use thereof.

Destroy this report when it is no longer needed. Do not return it to the originator.

Army Research Laboratory
Aberdeen Proving Ground, MD 21005-5067

ARL-TN-531 **April 2013**

Feature Extraction of High-Dimensional Structures for Exploratory Analytics

Andrew M. Neiderer
Computational and Information Sciences Directorate, ARL

REPORT DOCUMENTATION PAGE	Form Approved OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED (From - To)
April 2013	Final	May 2011–March 2012

4. TITLE AND SUBTITLE
Feature Extraction of High-Dimensional Structures for Exploratory Analytics

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)
Andrew M. Neiderer

5d. PROJECT NUMBER
2TEDUC

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
U.S. Army Research Laboratory
ATTN: RDRL-CII-C
Aberdeen Proving Ground, MD 21005-5067

8. PERFORMING ORGANIZATION REPORT NUMBER
ARL-TN-531

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES

14. ABSTRACT
This report summarizes lessons learned in a study conducted at the U.S. Army Research Laboratory (ARL) for visual examination of high-dimensional data (HDD). The initial effort included feature extraction (FE), as opposed to feature selection, of HDD structures for display. FE to a two- or three-dimensional Euclidean space allows exploration of underlying structure of HDD, even though the meaning of the actual data is obscured (latent variables). Some discussion of the FEs considered is given; further details can be found at URLs provided. Application of visual analytics technology (interaction/navigation within visualization) allows additional knowledge discovery. Research continues at ARL for the development of a method to gain insight into HDD, particularly in the application of an analytic strategy to terrorist data.

15. SUBJECT TERMS
dimensionality reduction, feature extraction, high-dimensional data, t-distributed stochastic neighbor embedding, neighbor retrieval visualizer, visual analytics

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Andrew M. Neiderer
Unclassified	Unclassified	Unclassified	UU	14	19b. TELEPHONE NUMBER (*Include area code*) 410-278-3203

Contents

List of Figures

1. Introduction

An attempt at interpretating a high-dimensional data (HDD) structure* typically necessitates reduction by (1) feature extraction (FE) and/or (2) feature selection (FS) of that most predictive for a given problem. Usually, humans best understand HDD when it is presented graphically as two-dimensional (2-D)/three-dimensional (3-D) object(s) (*1*). For more than three-dimensions, the human visual system/brain combination usually becomes less effective quite quickly (an estimation of how we actually reconstruct the third dimension was done at the Max Planck Campus in Tübingen, Germany, January 2012) (*2*). Our brain is amazingly good at recognizing objects under a large number of variations (*3*). Applying an FE irreversibly transforms data semantics, but the underlying topology can be further examined. On the other hand, an FS approach for display preserves the meaning of selected data, but plotting a subset may be misrepresentative of the HDD. Further research of FEs, FSs, and possibly a hybrid FE/FS for dimensionality reduction (DR) continues for optimal data visualization in the knowledge discovery process. Here, however, we report only on those FEs, both linear and nonlinear, that have been considered.†

Visualization/visual analytics (VA) for a structure in HDD space (d) involves re-embedding to a lower-dimensional space, i.e., a 2-D/3-D Euclidean space. The goal of a DR approximation is to preserve the structure as much as possible when mapping from d to 2-D/3-D. The result should be representative of the original data so there is no loss of information. The intrinsic dimension (P) is the minimum number of parameters necessary to account for observed properties in the original data ($P < d$) and reveals topological structure. This is a requirement for the re-embedding: topological properties must be preserved when going from d to D. Ideally, $D = P$.

Topology is an area of mathematics where the concern is not representing an object (or structure) in space but the connectivity, which must not be altered. In other words, twisting, deforming, and/or stretching are allowed but not tearing. As an example, a circle in 2-D is topologically equivalent to an ellipse.

In general, DRs try to eliminate any redundancy that may exist when projecting from d to 2-D/3-D. In figure 1, the first two approximations, principal component analysis and classical metric multidimensional scaling, are a linear DR (LDR). An LDR is based on a linear combination of

*The data is assumed to lie near or on a Riemannian manifold and is treated as a kind of object in space.

†Note that only FE is addressed here and, thus, will be treated as synonymous with DR unless stated otherwise. DR is usually defined as both FE and FS (see http://dictionary.sensagent.com).

the feature data. LDRs keep similar data points close together (distance preserving) when mapping from d to D. However, they cannot find curved manifolds* since they are based on a Euclidean distance.

A nonlinear DR (NLDR) approximation, which is also called a manifold* learner, preserves geodesic distances along the manifold—linear or nonlinear (see figure 1 for a comparison between Euclidean and geodesic distance). NLDRs include nonmetric multidimensional scaling (MDS), Isomap, locally linear embedding, Laplacian eigenmaps, stochastic neighbor embedding /t-distributed SNE (SNE/t-SNE), and neighbor retrieval visualizer/t-distributed NeRV (NeRV/t-NeRV) (see figure 2).† Most papers for an NLDR approximation demonstrate the algorithm using an artificial dataset, such as the Swiss roll or S-curve, and not for real-world data (*4*).

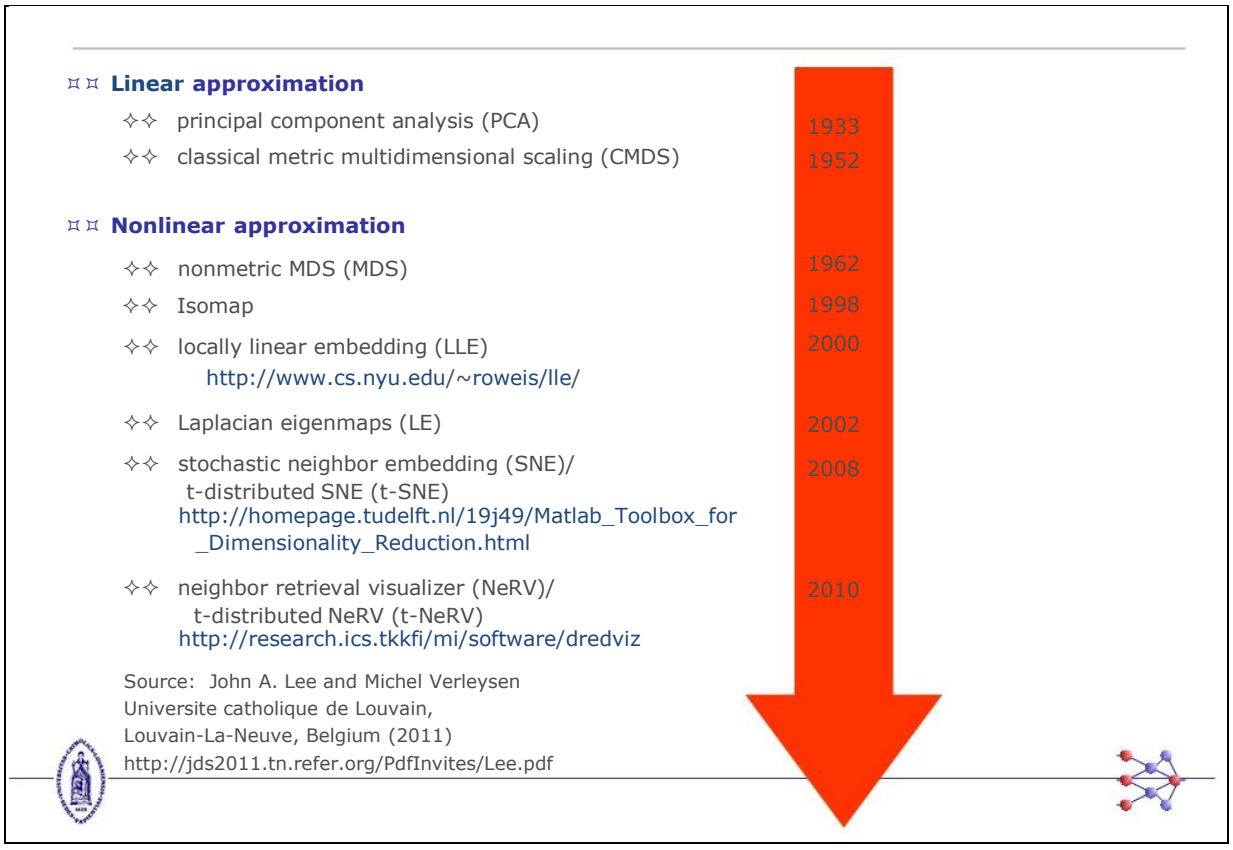

Figure 1. Some FEs for HDD.

*A manifold is locally Euclidean but may be globally curved. For example, the Earth is spherical in shape (global) but appears to be flat (local) to the human eye.

†A student-t distribution variant, in general, tries to create separation between natural clusters to alleviate the crowding problem.

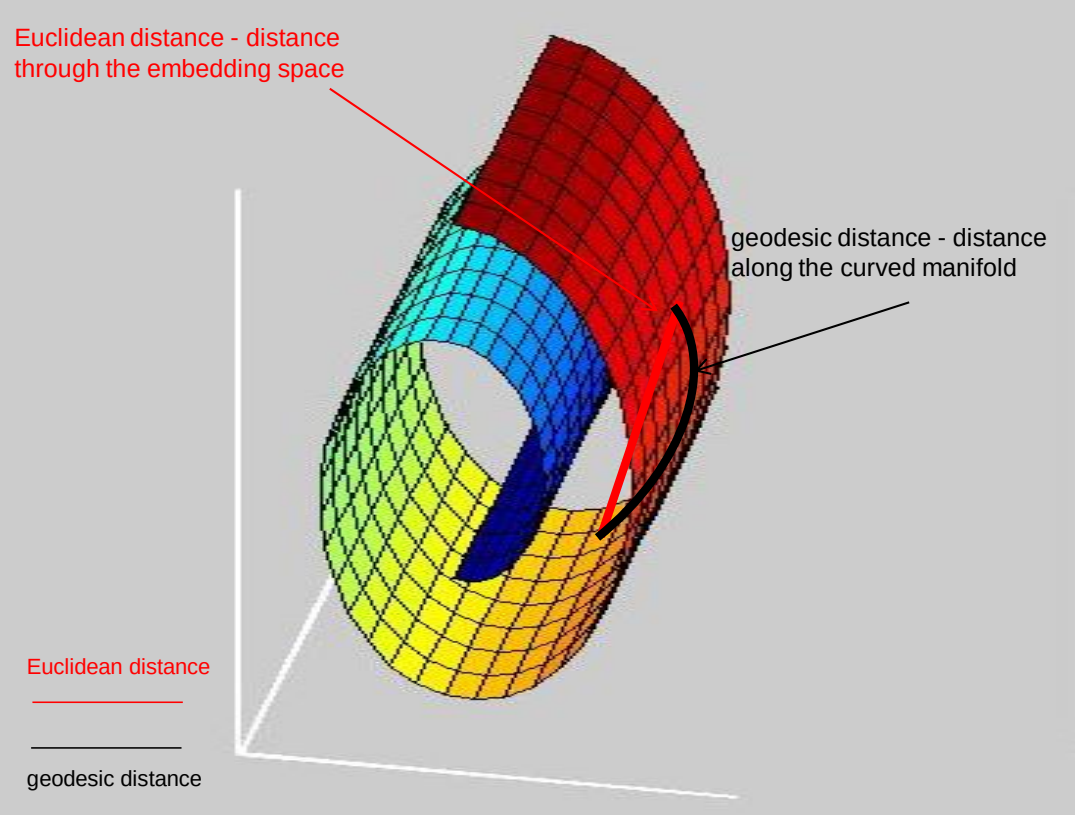

Figure 2. Comparison of Euclidean vs. geodesic distance. LDRs use metric based on the Euclidean distance between two points, while the NLDRs are based on geodesic distance. An NLDR successfully unrolls the curved manifold, whereas an LDR fails.

That same research paper (*5*) suggests that manifold learners may not be the best DRs for data visualization. The last two methods for NLDR in figure 1, namely SNE/t-SNE and NeRV /t-NeRV, are NLDRs that have been used with real-world data and are specifically designed for data visualization. Real-world data is usually highly curved but can be described by a reduced number of features.

2. Feature Extraction for Data Visualization

Recent research of FE for visualizing HDD structures states that this differs from applying previous manifold learners (or NLDRs) (*1*). Manifold learning, in most cases, was successfully done for an artificial dataset such as the Swiss roll. But further work by van der Maaten (*3*, *6*) at Tilburg University resulted in a new technique, called t-distributed stochastic neighbor embedding (t-SNE), for visualization of real-world HDD. In particular, van der Maaten used five real datasets in his dissertation: (1) MNIST hand-written digits (*7*); (2) Olivetti face data

(*8*); (3) the COIL-20 dataset; (4) word-features dataset; and (5) a Netflix dataset.* Although the manifold learners are invaluable conceptually, they are not the best for data visualization/VA.

In fact, a technical paper (*4*) in late 2011 from Aalto University's School of Science states that DR for data visualization is different from learning manifolds. The resulting NeRV/t-NeRV software includes a localized MDS.† Also note that SNE/t-SNE are special cases of NeRV when $\lambda = 1$ (see the equation in reference *4*).

3. Conclusions and Future Work

Research continues at the U.S. Army Research Laboratory for a means of using intelligence and HDD to gain insight in the knowledge discovery process and prevent terrorist activity from occurring. The distribution of data is not necessarily a geometrical locus but close to some manifold. Although FE application of a DR transforms the data semantics, the underlying topology can be further examined. The particular FE we are considering is NeRV/t-NeRV, which is a DR specifically for data visualization.

DR for data visualization can be achieved in one of two ways: FS or FE. Currently, we are researching fuzzy set theory (FST) and rough set theory (RST). Possibly, a hybrid FST/RST and NeRV/t-NeRV is being considered. Any continuous computation must be transformed to a discrete space for raster display.

*Note that van der Maaten presents the MATLAB implementations for many of the DRs (see http://homepage.tudelft.nl /19j49/Matlab_Toolbox_for_Dimensionality_Reduction.html).

†The dimensionality reduction software for information visualization from Aalto University is available at http://research.ics .tkk.fi/mi/software/dredviz/.

4. References

1. Lee, J. A.; Verleysen, M. *Nonlinear Dimensionality Reduction*; Springer Science+Business Media: New York, 2007.
2. Fooling Visual Neutrons Provides New Insight Into How the Brain Reconstructs the Third Dimension. http://tuebingen.mpg.de/en/homepage/detail/fooling-visual-neurons-provides-new-insight-into-how-the-brain-reconstructs-the-third-dimension.html (accessed 10 November 2011).
3. van der Maaten, L. Feature Extraction From Visual Data. Ph.D. Dissertation; Tilburg University, The Netherlands, June 2009.
4. Kaski, S.; Peltonen, J. Dimensionality Reduction for Data Visualization. *IEEE Signal Processing Mag.* **2011**, 100.
5. van der Maaten, L.; Hinton, G. Visualizing Data Using t-SNE. *Journal of Machine Learning Research* **2008**, *9*, 2579–2605.
6. van der Maaten, L. Learning a Parametric Embedding by Preserving Local Structure. Master's Thesis, TiCC, Tilburg University, LE Tilburg, The Netherlands, 2009.
7. MNIST database of handwritten digits, http://yann.lecun.com/excb/mnist/ (accessed 22 June 2011).
8. Olivetti Research, Ltd. face database, http://mambo.ucsc.edu/psl/olivetti.html (accessed 23 June 2011).

List of Symbols, Abbreviations, and Acronyms

2-D	two dimensional
3-D	three dimensional
d	data space
D	reduced Euclidean dimension
DR	dimensionality reduction
FE	feature extraction
FS	feature selection
FST	fuzzy set theory
HDD	high-dimensional data
LDR	linear dimensionality reduction
MDS	multidimensional scaling
MNIST	database of handwritten digits from NIST
NeRV	neighbor retrieval visualizer
NLDR	nonlinear dimensionality reduction
P	intrinsic dimension
RST	rough set theory
SNE	stochastic neighbor embedding
t-NeRV	t-distributed neighbor retrieval visualizer
t-SNE	t-distributed stochastic neighbor embedding
VA	visual analytics
Y	embedding space

NO. OF COPIES	ORGANIZATION
1 (PDF)	DEFENSE TECHNICAL INFORMATION CTR DTIC OCA 8725 JOHN J KINGMAN RD STE 0944 FORT BELVOIR VA 22060-6218
1 (PDF)	DIRECTOR US ARMY RESEARCH LAB RDRL CIO LL 2800 POWDER MILL RD ADELPHI MD 20783-1197
1 (PDF)	DIR USARL RDRL CII C A NEIDERER

INTENTIONALLY LEFT BLANK.